Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois, 60439

Appendix A

BlueGene/L Single Rack System Description

I. BlueGene/L Single Rack System

The hardware for a BlueGene/L System is organized at the highest level by rack. A rack contains the compute nodes, I/O nodes, the Torus network, the Global Tree network, the Global Interrupt (Barrier) network, and all of the cabling, power, and internal infrastructure required for operation. All of the hardware listed above is custom manufactured by IBM for the ANL BG/L Systems. In addition, the following off-the-shelf equipment is required for an operational system: a Service node, Front-end nodes, storage and an Ethernet network. All nodes with the exception of the Compute nodes will run Linux.

Compute Node (1,024 – 2048 processors)

The compute nodes contain two PowerPC 440 processors with 512 MB of RAM and will run a lightweight kernel to execute user-mode applications only. The Ethernet interface on the Compute nodes is not used; data is moved to and from the I/O nodes over the Global Tree network.

(2) 700 MHz* PPC440 CPUs
(512) MB RAM (embedded)
(6) connections to the Torus network @ 1.4Gb / link
(3) connections to the Global Tree network @ 2.8Gb / link
(4) connections to the Global Interrupt network
(1) connection to the Control network @ 100 Mb (JTAG)

I/O Nodes (32)

The I/O nodes run a full Linux kernel and act as a gateway for the compute nodes in their rack to the outside world. The I/O nodes present a standard Linux operating environment to the user. The Gigabit Ethernet interface of the I/O nodes will be connected into the core Ethernet switch.

(2) 700 MHz* PPC440 CPUs
(512) MB RAM (embedded)
(1) GigE adapter connected to the Ethernet network
(6) connections to the Torus network @ 1.4 Gb/link

(3) connections to the Global Tree network @ 2.8 Gb/link
(4) connections to the Global Interrupt network
(1) connection to the Control network @ 100 Mb (JTAG)

*Target nominal frequency

Service Node (1)

The Service node performs many functions for the operation of the BlueGene/L system, including system boot, machine partitioning, system performance measurements, and monitoring system health. The Service node will use DB2 as the data repository for system and state information.

eServer pSeries 655

(1) 1.7 GHz 4-way POWER4 processor (2 CPU cores)
(16) GB RAM
(2) 146.8 GB SCSI HDD
(4) GigE adapters

HMC (1)

The IBM Hardware Management Console for pSeries provides a standard user interface for configuring and operating partitioned and SMP systems. The HMC supports the system with features that enable a system administrator to manage configuration and operation of partitions in a system, as well as to monitor the system for hardware problems. It consists of a 32-bit Intel-based desktop PC with a DVD-RAM drive.

Front-End Nodes (4)

The Front-End nodes will be the direct interface to the users. Users will login, compile their applications, and submit their jobs to run from these nodes. They will have direct connections to both the BlueGene/L internal VLAN and the public Ethernet networks.

JS20

(2) 1.6 GHz PPC970 processors
(4) GB RAM
(2) 40 GB IDE HDD
(2) Integrated GigE adapters

Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois, 60439

Storage Nodes (16)

The Storage nodes provide mass storage for the BlueGene/L system. The Storage nodes each contain a ServeRAID 6i+ SCSI RAID Controller, which connects to six internal 146.8 GB 10K SCSI HDDs for a total of 880 GB raw storage per server or 14 TB total raw storage (11.7 TB usable). Each storage node includes two integrated Broadcom GigE adapters, which are connected to the core Ethernet switch for a total of 32 GigE connections to the storage. This matches the 32 GigE connections from the I/O nodes to the switch.

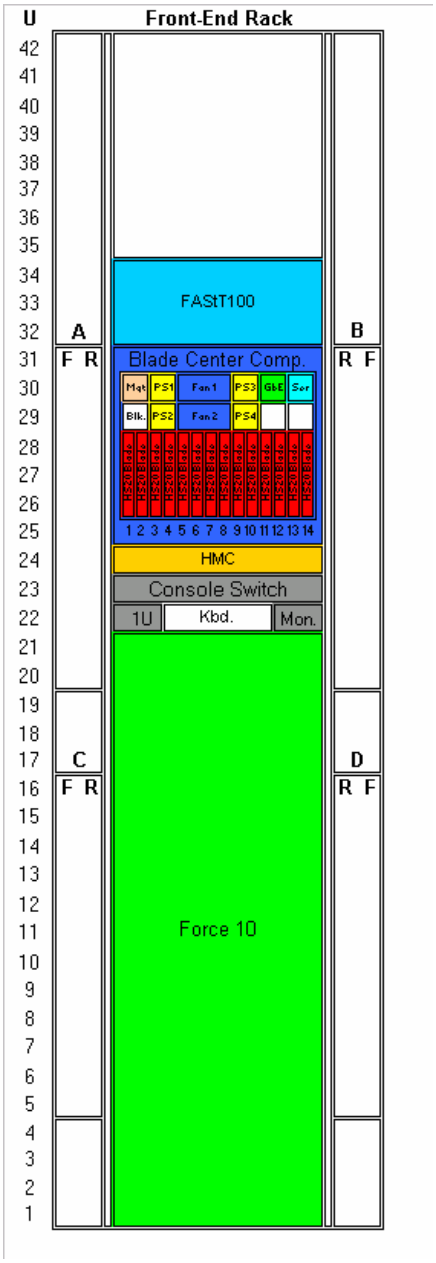
xSeries 346

(2) 3.4 GHz Xeon processors
(4) GB RAM
(2) Integrated GigE adapters
(1) ServeRAID 6i+ SCSI RAID controller
(6) 146.8 GB 10K SCSI HDDs

II. BlueGene/L System Racks

In addition to the BlusGene/L System and Service node, which each require their own rack the following two racks are required:

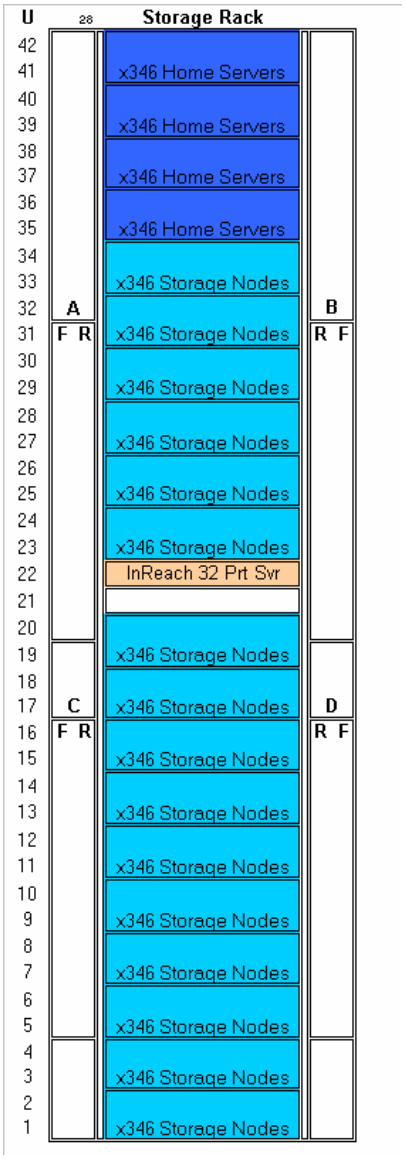
Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois, 60439



The Front-End Rack contains:

(1) Rack
(1) Terminal Server
(1) Keyboard & Monitor Server
(4) JS20 (as part of a BladeCenter)
(1) FAStT100 (now named DS4100)
(1) HMC
(1) Force10 Switch (customer supplied) with 48-port 10/100/1000 Ethernet modules

Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois, 60439



The Storage Rack contains:

(1) Rack
(1) Terminal Server
(4) Home File Servers
(16) Storage Nodes

III. Networks

There are five independent networks in a BlueGene/L system:

- **Three-Dimensional Torus** – point-to-point

The Torus network is used for general-purpose, point-to-point message passing and multicast operations to a selected “class” of nodes. The topology is a three-dimensional torus constructed with point-to-point, serial links between routers embedded within the BlueGene/L ASICs. Therefore, each ASIC has six nearest-

neighbor connections, some of which may traverse relatively long cables. The target hardware bandwidth for each Torus link is 175 MB/s in each direction for link for a total of 2.1 GB/s bidirectional bandwidth per node.

- **Global Tree** – global operations

The Global Tree network is a high-bandwidth one-to-all network that is used for broadcast traffic and also to move process and application data from the I/O nodes to the compute nodes. Each compute and I/O node has three links to the global tree network at 350 MB/s per direction for a total of 2.1 GB/s bidirectional bandwidth per node. Latency on the Global Tree network is less than 2.5 μs from the bottom to top of the tree, with an additional 2.5 μs latency to broadcast to all.

- **Global Interrupt** – low-latency barriers and interrupts

The Global Interrupt network is a separate set of wires based on asynchronous logic that form another tree that enables fast signaling of global interrupts and barriers (global AND or OR). The target latency to perform a global barrier over this network for a 64K node partition is approximately 1.5 microseconds.

- **Gigabit Ethernet** – File I/O and host interface

The Gigabit Ethernet network consists of all the I/O nodes and discrete nodes connected to a standard Gigabit Ethernet switch. A customer-supplied Gigabit Ethernet switch with at least 96 non-blocking GigE ports will be provided. The Compute nodes are not directly connected to this network; all traffic is passed from the Compute node over the Global Tree network to the I/O node and then onto the Gigabit Ethernet network. Note that any network communication requirements outside the scope of the BlueGene/L system will be the responsibility of ANL.

- **Control Network** – boot, monitoring, and diagnostics

The Control network consists of a JTAG interface to a 100 Mb Ethernet interface with direct access to shared SRAM in every Compute and I/O node. The Control network is used for system boot, debug, and monitoring. It allows the Service node to provide runtime noninvasive RAS support as well as noninvasive access to performance counters.

Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois, 60439

IV. System Configuration Details

V22	IBM eServer	Qty
	Processor Hardware - BG/L Core Rack	
XXXXXXX	1024 Dual-CPU PPC 440 700Mhz 512MB RAM 16 I/O Nodes	1
XXXXXXX	Additional I/O nodes	16
	Processor Hardware - Service Server	
7039-651	pSeries 655 Rack Server	1
3275	146.8GB 10K RPM	2
3800	Power Cable Group	1
4457	8GB memory card	2
4609	drawer place indicator	1
4651	rack indicator, rack 1	1
4961	UNIV 4-P 10/100 ETHERNET ADAP	1
5518	4-way 1.7Ghz power 4 processor	1
5706	2 port PCI-X Adpt	1
6199	DC power converter assembly	1
6430	service support processor	1
6569	U320 SCSI backplane	1
6591	Dual CEC side by side mount	1
8121	Atach cable to host, 15M	1
9300	Language group US/English	1
9800	Power Cord US/Canada	1
7315-CR2	HMC 1:RACK-MOUNTED HMC	1
960	HMC FOR PWR4 LIC.MACH.CODE	1
966	PSERIES INDICATOR	1
2943	8-PORT ASYN ADP EIA-232/RS-422	1
2944	128-PORT ASYN CONTR, PCI BUS	1
4242	6 FOOT EXTENDER CABLE DISPLAYS	1
4651	RACK INDICATOR,RACK #1	1
8121	ATTCH CAB, TO HOST, 15 METER	1
8131	128-PORT ASYNC CONTR CAB, 4.5M	1
8133	RJ-45 TO DB-25 CONV. CAB.	4
8136	RACK MOUN.REM.ASYN.NODE 16PORT	1
8800	QUIET TOUCH KEYB-USB, US ENG.	1

8841	MOUSE- W/ KEYB ATTCH CAB	1
9300	LANG GRP SPEC.-US ENG.	1
9800	PWR CORD SPEC.- US/CANADA	1
9911	PWR.CORD(4M)SPCF-ALL	1
7040-W42	Rack	1
0211	Rack Cont Spec: Half Drawer	1
3754	Service Toolkit	1
4651	Rack Indicator	1
6076	Front door	1
6078	Rear Door	1
6186	Bulk Power Regulator	2
6187	Power Contr: 4 Cooling Fans	2
8123	Attach Cable , HW Mgmt 15M	1
8688	14 Ft Line Cord	2
8690	Bulk Power Ass., Redundant	1
9300	Language Group - English	1
	Processor Hardware - BladeCenter Front-End Servers	
86772XX	eServerBladeCenter7UChassis,14bays,2hot-swap&redundantswitch&load-bal.1800Wpwrsupmodules	1
884221X	JS20 2x1.6 512/512/0GB	4
73P2276	1GB PC2700 CL2.5 ECC DDR SDRAM RDIMM	16
48P7063	IBM eServer BladeCenter (TM) 40GB 5400-rpm ATA-100 HDD	8
13N0568	BladeCenter4-portGigabitEthernetSwitchModule	2
13N0570	BladeCenter1800wPowerSupplyModules(2)	1
	Processor Hardware -Storage Server	
884022U	xSeries 346 3.20G 1 1/0	16
73P2866	2GB PC2-3200 (2x1GB) ECC DDR2 SDRAM RDIM	32
26K5097	xSeries 625W Hot Swap Redundant Power Supply (N. America)	16
71P8642	ServeRAID-7k Controller	16
32P0728	e1350 146.8GB, 10K RPM, Ultra320 SCSI Hot-swap HDD	96
58P8665	e1350 Option Install Fee - Devices with 4 or More Options	16
06P7515	e1350 Install >1U Device	16
	Processor Hardware - Home File Server (FS)	
884022U	xSeries 346 3.20G 1 1/0	4
73P2866	2GB PC2-3200 (2x1GB) ECC DDR2 SDRAM RDIM	16

32P0728	e1350 146.8GB, 10K RPM, Ultra320 SCSI Hot-swap HDD	8
24P0960	; e1350 FC2-133 PCI 2GB/sec Fibre Host Bus Adapter - 64-bit, 1/2 length, low-profile	4
26K5097	xSeries 625W Hot Swap Redundant Power Supply (N. America)	4
58P8665	e1350 Option Install Fee - Devices with 4 or More Options	4
	Storage Controller - DS4100 Gbit Home FS	
1724100	IBM TotalStorage DS4100 ;;;; (formerly IBM TotalStorage FAStT100 Storage Server)	1
24P8068	e1350 ; FAStT600 Linux/Intel Host kit	1
19K1271	e1350 ; FAStT600 Short-wave SFP Module ;	4
90P1350	SATA 250GB 7200rpm Disk Drive Module	14
58P8665	e1350 Option Install Fee - Devices with 4 or More Options	1
06P7515	e1350 Install >1U Device	1
	Network Hardware - Terminal Server 3rd Party	
02R2239	e1350 InReach LX-40325 32 Port Console Server	1
31P4310	e1350 Install 1U Device	1
24P7218	e1350 1U OEM Common Rack Mounting Hardware	1
	Cabinet Hardware - xSeries Main 42U	
141042X	e1350 IBM eServer Cluster 42U Rack ;	1
32P1766	e1350 DPI Single-phase Front-end Power Distribution Unit	2
32P1736	e1350 DPI 100-240V Power Distribution Unit	6
94G7448	4.3m IEC Power Cable C13/ C14	21
17231NX	e1350 1U 17" Console Kit without keyboard	1
73P3144	Travel Keyboard (US English)	1
1735L04	e1350 NetBAY Local Console Manager - 4-port Switch w/ RJ45 connectors	1
32P1652	e1350 Cable Kit, KVM Conversion (KCO) Long ;; 1.5m	5
24P7218	e1350 1U OEM Common Rack Mounting Hardware	1
02R2271	e1350 Multi-Colored Ethernet Cable Kit (qty derived from MANU-xxM items for pricing - qty is NOT a cable count)	33
MANU-L3M	e1350 CAT5E Cables, Intra Rack, =< 3m. (IBM internal item for mfg use only)	25
MANU-L10M	e1350 CAT5E Cables, Inter Rack, = 10m. (IBM internal item for mfg use only)	4
31P6327	e1350 Fibre Channel 2GB Cable, LC to LC 3m ;;; (PCI	8

	to FAStT600, T700, T900 & FASt to EXP in same rack)	
24P7963	e1350 Individual Serial, RJ45 to DB9, 12' ;;;; (Mgmnt Node to Terminal Server)	1
06P7514	e1350 Enterprise 42U Rack Prep Fee	1
58P8608	e1350 Cluster System Validation and Test - 1st 42U Rack	1
24P7955	Ship Group (includes publications and possible accessories)	1
	Cabinet Hardware - xSeries Expansion 42U	
930842S	e1350 IBM eServer Cluster 42U Rack ;	1
32P1766	e1350 DPI Single-phase Front-end Power Distribution Unit	2
32P1736	e1350 DPI 100-240V Power Distribution Unit	6
31P6326	e1350 Fibre Channel 2GB Cable, LC to LC 10m ;; (PCI to FAStT700 up to 6M (8 racks) away)	8
06P7514	e1350 Enterprise Rack Prep Fee	1
58P8609	e1350 Cluster SW Install and Test - Subsequent Rack	1
	Software - IBM Supplied	
5733-BG1	BGL CORE: Compute Node Kernel (CNK), MPI support for hw implementation and Abstract Device Interface, Core Monitoring & Control System (CMCS), System Diagnostics, and External Scheduler for LoadLeveller	1
	Mini-Control Program (for I/O Nodes)	32
	GNU Toolchain Patches for CNK (BGL changes to support GNU)	4
5639-LNX	Service Node OS: SUSE® LINUX Enterprise Server 8 for i/p Series	1 (<16 CPU)
D54KWLL	XLF (Fortran) for Linux	4
	XLF Runtime for BG/L	4
D54L0LL	XL C/C++ Advanced Edition for Linux	4
	XL C/C++ Runtime for BG/L	4
D518GLL	DB2 UDB Enterprise Server Edition	4
	DB2 UDB Enterprise Server Edition - Client Code	16
	Software - ANL Supplied	
MPICH2	MPI (MPICH2) Library	4
SLES9	Front End Node OS: SUSE® LINUX Enterprise Server 9 on IBM POWER	24 (<16 CPU)
JRE	Java Runtime JRE for SUSE	20
	GNU Toolchain (glibc, gcc, binutils, gdb)	4
	Python	tbd